Roles and Reusability of Video Data in Social Studies of Interaction

SCARP Case Study No. 5

Summary and Recommendations

Angus Whyte

Digital Curation Centre, University of Edinburgh

Executive Summary

Social science researchers are making increasing use of digital video. All of us, researchers or not, have an alluring range of commercial web sites for sharing video, although these do not cater for long-term reuse of video in research. But what kind of roles does video fulfil as research data? And what curation issues and challenges does video raise for researchers and their institutions? The phenomenal growth in public use of digital video is a topic of social research; in the first six months of 2008, users of Youtube uploaded more video footage than the top three U.S. TV networks would have broadcast if they had been operating 24 hours per day over their sixty-year lifespan (Wesch, 2008). Yet there have been few studies of social scientists' own uses of digital video data in their research.

Video data is rich in potential for repeated study within and between projects. Aiming to contribute to developing that potential, the report explores several fields of social research where video corpora are being developed, and possibilities for secondary analysis are being actively explored. Despite recent e-Social Science work in tools for analysis of multi-modal corpora, previous studies of qualitative data reuse in social science have given little attention to this area.

An increasing range of social science and technology related fields are using corpus-based approaches[1], describing and analysing patterns in and across examples of human activity, recorded in text, sound, still and moving image, and as traces of digital interaction. Video is at the core of multimodal corpora, a backing track or temporal map, allowing different views and time-based data sources to be overlaid; facilitating inter-disciplinary reuse.

The growing inter-disciplinary use, complexity and size of video data make it important for research data services to understand and support it. The report uses the DCC Curation Lifecycle Model to identify shared needs for support with curation. It also highlights disciplinary differences within the relatively small area of interaction research, pointing to diversity in the roles of video and the contextual information needed to reuse it.

The study focuses on two main groups of researchers; SEDIT (Scottish Ethnomethodology, Discourse and Interaction) is an informal cross-disciplinary network of researchers, some of whom use video as observational data for ethnographic analysis in Human Geography and in Computing. The second group use video as experimental data in studies of eye movements and scene perception, based in the Visual Cognition Research Group of the University of Edinburgh Dept of Psychology.

Chapter 1 profiles the researchers, projects and groups that were the focus of the study. Data for the case study was obtained from interviews with members of both groups, and from participant observation in SEDIT 'data sessions' where video data are collaboratively analysed. The chapter includes a literature and landscape review of developments in the fields mainly concerned, providing background on the relevance of video and the rationale for sharing data. Among these drivers for video data curation, the chapter describes e-Social Science projects addressing researchers' needs for real-time data sharing and analysis tools.

Chapter 2 presents three main themes that arose from the case study. Interviews and observations were used to identify challenges that video data poses for a 'lifecycle management' approach to planning data curation. The first theme is the *diversity of research practices* involved. Video serves different roles across and within research fields, and at different stages of a project. It may be publicly accessed and used at the beginning of its life as data (e.g. as web video clips), or made public at the end of the lifecycle (e.g. clips on researchers' websites), and shared at various points in-between (e.g. at 'data sessions' and conferences). Researchers with broadly similar analytic orientations, e.g. to ethnographic

[1] This definition is meant to include any methodology where collected 'examples of human activity' are produced as a resource for repeated analysis.

observation, use video material differently according to the research topic or application area. Technology choices were more closely related to researchers' communities of interest or practice than to wider disciplinary contours; especially as researchers import methodologies from disciplines other than that which their research group is institutionally aligned with.

The second theme is the *uncertainty* affecting data curation planning decisions based on the Curation Lifecycle model, given rapid changes in technology, the complexity of the format choices to be made, and the need in exploratory research to begin with open questions about the data to be acquired and analysed. Methodological preferences and research topics influence the level of image detail needed to describe, analyse and interpret video alongside other data associated with it, depending for example on the required attention to the detail of gestures, or shifts in gaze.

Methodology similarly affects *appraisal and selection* of video data. In ethnographic studies data quality judgements involve trade-offs between the unfolding relevance of the material, the audio and visual clarity, and legal or ethical factors. Similar factors affect experimental psychology where video is used as an experimental stimulus. Here the need to maximise image clarity is driven by the need to perform statistical analysis on frame-by-frame changes in the image content, correlating these with other experimental variables. Studies in this field are 'data driven', using hypotheses based on exploring the patterns found rather than on theoretical models. Finding moving images to use as experimental stimuli that are controllable, ecologically valid, and are likely to yield informative results involves trial and error with a variety of content genres. Meanwhile the eye movement research community is only beginning to formulate expectations of how results should be made available.

Researchers generally did not plan for long-term *preservation* given the uncertainties of post-project funding and the confidentiality, consent and copyright issues in sharing video openly. In both observational and experimental studies, how much of acquired video data can be shared beyond the research team is a legal/ethical question as much as it is about the infrastructure for sharing video. Storage constraints affect all parts of the curation lifecycle for video data, each step entailing questions about 'where it will fit now' given the available and affordable capacity, and 'where it could go later' given the ethical limitations on disclosure.

The *third* theme of Chapter 2 is the nature of the *context information* needed to reuse archived video objects. Much social research that uses video is concerned with understanding phenomena in their natural setting. This makes it problematic to differentiate between data and context, especially if context is treated as a static description. Social research has different perspectives on what 'context' is, and whether it can ever be satisfactorily recorded. Video archives would be better enabled to address this by treating context information as a dynamic property, resulting from dialogue between the original researcher and reusers.

Chapter 3 considers the implications for curation lifecycle managing. To manage the range of possibilities and contingencies, a more iterative approach is proposed involving three main cycles of curation. Firstly a '*planning and piloting'* phase begins with the data management plan and then revises this in light of the data initially gathered. The main '*project curation*' phase begins with selection of data for analysis and implementation of tools and standards to enable involvement of colleagues and peers in that. Then the '*long term curation*' phase begins as researchers' work up the data for publication and longer-term preservation and reuse.

The report summarises curation strategies researchers in this study adopt, and provides sources of further guidance, drawing on the literature including recent landscape reviews for JISC and the AHRC. It also draws on discussions with University of Edinburgh research data service providers aiming to envisage how, in this and other UK institutions, the respective roles of research groups and centrally provided data archives may evolve to handle video material.

Conclusions and Recommendations

The richness of digital video data for repeatedly analysing human interaction is driving the development of shared data resources and tools in social research fields that are concerned with closely analysing language and interaction. These include a range of 'data driven' traditions that are finding novel ways to identify patterns in their data for further interpretation or experimentation. Just as technology has underpinned the development of corpus-based linguistics, development support for online corpora of video and related materials is likely to promote reuse of data in interaction-oriented social science, building on methodological traditions of reusing and sharing examples of interaction.

The curation needs of researchers in multimodal interaction differ in important ways from those of qualitative or mixed-method researchers in other areas. One of the main differences is in the contextual information requirements that would support reuse. In many areas of qualitative social research lack of access to the original research context is commonly seen as a major barrier to secondary analysis. In interaction analysis it is less of a barrier. Where the analysis of particular actions or behaviour depends on understanding their place in an unfolding sequence of interaction that has been audio-visually recorded, the data 'content' is itself part of 'the context'. Sharing this data depends partly on documented detail provided up-front by the data creator, but mostly on the possibilities that collaboration with others affords for developing a richer analysis of it.

The needs for collaboration support have begun to be addressed through ESRC funded work by the UK Data Archive, and in the UK e-Social Science programme. Meanwhile more advanced models for archiving and curating annotated video or multimodal[2] corpora are being developed and implemented by linguistic archives in the US, France and the Netherlands. These provide searchable corpora comprising video and synchronised data that may be browed with their annotations online, to aid and stimulate reuse. UK researchers, for example in the *DreSS*[3], and *AMI* [4]projects, are already adopting techniques to develop support for cross-disciplinary interaction analysis. Archival and metadata models from the linguistic community may also have wider influence on reuse in social interaction research.

Sustaining the accessibility and reusability of digital video-based research materials is a challenge to domain-based archiving initiatives and to national and institutional data services. The challenges include identifying and fulfilling the various roles that video may play at different stages in the research process, and enabling appropriate legal and ethical controls on data access.

Recommendation 1- DCC, JISC and other research funders should develop the e-infrastructure for multi-disciplinary interaction research by facilitating workshops to bring together the disciplines involved, disseminate relevant tools, and explore more effective ways to browse and annotate multimodal data in data repositories. This would take forward work piloted by UKDA and by the DreSS project.

When video constitutes a significant proportion of the data to be created or collected for research purposes, decisions on how to manage it are likely to be revisited repeatedly, long before any of it is archived for potential reuse. Early decisions on the data to be collected; options and formats for capture and storage, and the tools and resources for analysis are likely to change throughout a project. Methodology may require initial questions to be refined in light of patterns identified from early data collection, which may also narrow technical options. A phased approach to assessing the risks to re-usability is needed especially given

---

[2] The term 'multimodal' is favoured by linguistic and psychology researchers over 'multimedia' as it better reflects their view of audio-visual and instrumented data as recordings of 'modes of communication'.

[3] Digital Records for e-Social Science (DreSS) available at: http://web.mac.com/andy.crabtree/NCeSS_Digital_Records_Node/Welcome.html (August 2009)

[4] Augmented Multiparty Interaction (AMI) available at: http://corpus.amiproject.org/ (August, 2009)

the changing and complex relationships between data policies, ethics and rights issues.

Recommendation 2 –The DCC Curation Lifecycle Model is an 'ideal type' and rather than used as a one-off framework for Data Management Planning it should be used iteratively during research projects, by periodically reviewing the Data Management Plan so that research materials that have been collected can be used effectively by the core research team, collaborators and other potential reusers.

The case study illustrates some of the diversity of research practices in the social sciences, and their influences on the re-usability of video data. To identify relevant support for curation of this research material the report includes a landscape review of tools, resources, and advice services available to the UK Higher Education community and should interest researchers and service providers in this rapidly evolving area. The study also indicates that preservation and curation of video and multimodal research data would benefit if researchers had better-coordinated support for video, across local and national institutional services. Initiatives to publish video corpora are likely to be best led by researchers in the domains concerned, but with coordinated support from institutional and national data repositories in such key areas as storage management, format migration/ transcoding, metadata implementation, ethics and IPR – areas that may already be addressed by institutions' e-learning initiatives. The alternative is likely to see researchers increasingly using commercial web enterprises oriented to 'user generated' video content in ways that neither comply with legal and ethical obligations nor keep data accessible and reusable.

Recommendation 3 – DCC should collaborate with relevant Research Councils, JISC Digital Media and JISC Legal Information to guide institutions, research ethics committees, and researchers on planning and managing the curation of video and multimedia research data.

Recommendation 4 – HEI's should consult researchers on the methodological and technical issues affecting the reusability of video and multimodal data they would want to submit to institutional or subject data repositories, and coordinate the support they provide with the relevant services provided by JISC and other agencies.